

INF 558: Building Knowledge Graph

Homework 4: CRF

Q.1) What was the information you were looking for? Describe the labels you chose and why. Also include at least one screenshot of the webpage you are using and show where the information you are looking for.

My source of unstructured data is from indeed.com, which is a job portal. Also my previous assignments retrieved structural data and our project is based on scraping from job portals, hence this selection.

I have used scrapy to crawl the job posting and retrieve the unstructured job description.

Scrapy spider: indeed.py

I have stored the output of this crawled data in a json file: Abhishek_Dhameja_hw4.json.

The following are the key attributes I planned to retrieve from the unstructured data:

1. Company Name (C)
2. Position (P)
3. Skills (S)

The screenshot shows a job posting for Google. The title is "Software Engineer, Tools and Infrastructure". The location is "San Francisco, CA". The job description is divided into sections: "Position", "Company Name", and "Responsibilities". The "Position" section describes the role as a Software Engineer, Tools and Infrastructure, focusing on building software that empowers engineering teams. The "Company Name" section describes Google's mission and its focus on organizing the world's information. The "Responsibilities" section describes the role of a Software Engineer at Google, focusing on building software that empowers engineering teams. Annotations include orange boxes around the job title, location, and company name, and a red box around the "Position" section. Arrows point from the "Structured Data" label to the job title and location, and from the "Unstructured" label to the job description text.

Software Engineer, Tools and Infrastructure ← Structured Data

Google ★★★★★ 2,061 reviews - San Francisco, CA

A line of code can be many things - an amazing feature, a beautiful UI, a transformative algorithm. The faster this line of code reaches millions of users, the sooner it impacts their lives. As a **Software Engineer, Tools and Infrastructure**, you will be at the heart of **Google's** engineering process building software that empowers engineering teams to develop and deliver high quality products quickly. We are focused on solving the hardest, most interesting challenges of developing software at scale without sacrificing stability, quality, velocity or code health.

Position

We ensure **Google's** success by partnering with engineering teams and developing scalable tools and infrastructure that help engineers develop, test, debug and release software quickly. We impact thousands of Googlers and billions of users by increasing the pace of product development and ensuring our products are thoroughly tested. We are champions for code health, testability, maintainability and best practices for development and testing.

Company Name

Having access to all of **Google's** platforms and vast compute resources provides a unique opportunity to grow as an **engineer**. We typically work in small, nimble teams that collaborate on common problems across products and focus areas. As a result, the exposure to this broad set of problems provides diverse technical challenges as well as accelerated career growth.

Google is and always will be an engineering company. We hire people with a broad set of technical skills who are ready to take on some of technology's greatest challenges and make an impact on millions, if not billions, of users. At Google, engineers not only revolutionize search, they routinely work on massive scalability and storage solutions, large-scale applications and entirely new platforms for developers around the world. From AdWords to Chrome, Android to YouTube, Social to Local, Google engineers are changing the world one technological achievement after another.

Responsibilities

• Lead/contribute to engineering efforts from design to implementation, solving complex

Google

Follow Get job updates from Google

Google ★★★★★ 2,061 reviews

Google's mission is to organize the world's information and make it universally accessible and useful. In 1998, we started with two...

← Unstructured

Let employers find you

Thousands of employers search for candidates on Indeed

Upload Your Resume

Responsibilities

- Lead/contribute to engineering efforts from design to implementation, solving complex technical challenges around developer and engineering productivity and velocity
- Design and build advanced automated build, test and release infrastructure
- Drive adoption of best practices in code health, testing, and maintainability
- Analyze and decompose complex software systems and collaborate with cross-functional teams to influence design for testability

Skills

Qualifications Minimum qualifications:

- BS in Computer Science or related technical field or equivalent practical experience
- Software development experience in one or more general purpose programming languages
- Experience in at least one of the following: test automation, refactoring code, test-driven development, build infrastructure, optimizing software, debugging, building tools and testing frameworks

Preferred qualifications:

- Master's or PhD in Computer Science or related technical field
- Experience with one or more general purpose programming languages including but not limited to: Java, C/C++, C#, Objective C, Python, JavaScript, or Go
- Scripting skills in Python, Perl, Shell or another common language

Q.2) What kind of tags did you tag your data with? Explain your choices.

Following tags are collected from the Job description which are scraped from the website:

1. Company Name (tagged as "C")
2. Position (tagged as "P")
3. Skill (tagged as "S")
4. Irrelevant data (tagged as "I") (the data which do not belong to any of the above three categories)

I have manually marked 50 training records and 20 test records with C/P/S tag.

Build the CRF model on the train data.

Predicted and compared with the manually tagged test data.

Q.3) Report your classifier's precision, recall and F-1 measure. Why did your classifier perform well (or not satisfactorily)?

Find the snapshot of the classifier's performance :

```
C:\Python27\python.exe E:/Workspaces/Python/inf558/Abhishek_Dhameja_hw4/Abhishek_Dhameja_crf.py
      precision    recall  fl-score   support
C         1.00      0.09      0.17         65
I         0.97      1.00      0.98      8361
P         0.52      0.38      0.44         29
S         0.84      0.28      0.42        304

avg / total         0.96      0.96      0.95      8759

Process finished with exit code 0
```

1. Classifier performs decently in case of the P and the S tag, as all the jobs did mention in the description the position and the skill.
2. It performs poorly in case of the C tag as most of the company have not mentioned their names in the description(as already present in the structured heading)(We can retrieve that information from the heading)
3. The classifier's performance is boosted as I have used POS tags as one of the features among others